

Effective Resistances, Statistical Leverage, and Applications to Linear Equation Solving

Petros Drineas *

Michael W. Mahoney †

Abstract

Recent work in theoretical computer science and scientific computing has focused on nearly-linear-time algorithms for solving systems of linear equations. While introducing several novel theoretical perspectives, this work has yet to lead to practical algorithms. In an effort to bridge this gap, we describe in this paper two related results. Our first and main result is a simple algorithm to approximate the solution to a set of linear equations defined by a Laplacian (for a graph G with n nodes and $m \leq n^2$ edges) constraint matrix. The algorithm is a non-recursive algorithm; even though it runs in $O(n^2 \cdot \text{polylog}(n))$ time rather than $O(m \cdot \text{polylog}(n))$ time (given an oracle for the so-called statistical leverage scores), it is extremely simple; and it can be used to compute an approximate solution with a direct solver. In light of this result, our second result is a straightforward connection between the concept of graph resistance (which has proven useful in recent algorithms for linear equation solvers) and the concept of statistical leverage (which has proven useful in numerically-implementable randomized algorithms for large matrix problems and which has a natural data-analytic interpretation).

1 Introduction

The problem of approximating the solution to a set of linear equations defined by a Laplacian constraint matrix has been of interest recently due to a series of remarkable papers by Spielman and Teng [35, 37, 36]. (This work builds on ideas originally introduced by Vaidya and developed by others [9, 8, 7].¹) While introducing several novel theoretical perspectives, this work on “nearly-linear-time” algorithms has yet to lead to practical algorithms. In this paper, we describe two related results in an effort to bridge this theory-practice gap.

Our first and main result, to be described in Section 2, is a simple algorithm for computing an approximate solution to a set of linear equations defined by a Laplacian constraint matrix. The simplicity of the algorithm permits us to identify a simple connection (that to our knowledge

*Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, drinep@cs.rpi.edu.

†Department of Mathematics, Stanford University, Stanford, CA, mmahoney@cs.stanford.edu.

¹Briefly, recall that to solve a system of linear equations, $Ax = b$, one can use either direct methods or iterative methods [26, 39]. Iterative methods, such as Chebyshev or Conjugate Gradients, compute successively better approximations to x by performing successive matrix-vector multiplications. The number of iterations typically depends on the condition number $\kappa(A)$ of A , where $\kappa(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$ is the ratio of the extreme (nontrivial) eigenvalues of A , via a multiplicative factor of $\sqrt{\kappa(A)}$. Preconditioning refers to a class of methods to solve $B^{-1}Ax = B^{-1}b$, where the preconditioning matrix B is chosen such that $\kappa(B^{-1}A)$ is small and such that it is easy to solve for $Bz = c$. Vaidya introduced the idea of using combinatorial methods to precondition Laplacians of graphs with Laplacians of their subgraphs. It is known that if one wants to precondition any symmetric diagonally dominant matrix, then it suffices to find a preconditioner for a related Laplacian matrix [7]; and, moreover, that preconditioning matrices that arise in many applications can be reduced to the problem of preconditioning diagonally dominant matrices [10]. Vaidya’s methods have been extended [7, 9, 8], and they were used by Spielman and Teng to approximate the solution to diagonally dominant linear systems in time that is “nearly-linear” in the number of nonzero entries in their defining matrices [36].

has been overlooked) with other recent work in the theory and (numerical and data) application of randomized algorithms for matrix problems. Thus, our second result, to be described in Section 3, is to identify and discuss the connection between the concept of *statistical leverage* and the concept of *graph resistance*. The latter concept has a long history in spectral graph theory [16], and recently it has proven useful in algorithms for linear equation solvers [34, 3]. The former concept also has a long history, but in statistics and diagnostic data analysis [15]. Moreover, recently, it has been demonstrated to be the key structural quantity to understand in order to bridge the theory-practice gap between theoretical work on randomized algorithms for large matrices and applications (both numerical-implementation and data-analysis applications) of this “randomized matrix algorithm” paradigm [28, 27, 24, 30, 2].

1.1 Laplacian matrices

Consider a graph $G = (V, E)$ with n vertices and m weighted, undirected edges. We will assume that all the weights are positive. Then, we can construct the so-called Laplacian matrix $L \in \mathbb{R}^{n \times n}$ of G . Let $w_{ij} \geq 0$ denote the weight of the edge joining vertices i and j ; clearly $w_{ij} = 0$ if no such edge exists. In the most common definition of the Laplacian matrix L , the off-diagonal entries of L (L_{ij} , $i \neq j$) are set to $-w_{ij}$, while the diagonal entries L_{ii} (for all $i = 1, \dots, n$) are equal to the “weighted degree” of vertex i , i.e., $L_{ii} = \sum_{j=1}^n w_{ij}$. By definition, L is a symmetric matrix of rank at most $n - 1$, since the all-ones vector is clearly in the null space of L .

A somewhat less common definition of the Laplacian matrix follows from the so-called edge-incidence matrix of the graph G . Let $B \in \mathbb{R}^{m \times n}$ denote the edge-incidence matrix of the undirected graph G , constructed as follows: each row of B corresponds to an edge of G ; assuming that an (arbitrarily-oriented) edge of G starts at vertex i and ends at vertex j , the i -th entry in the corresponding row of B is set to $+1$, the j -th entry is set to -1 , and the remaining entries are all set to 0. Thus, B has two non-zero entries per row for a total of $2m$ non-zero entries. Also, let $W \in \mathbb{R}^{m \times m}$ be a diagonal matrix containing the edge weights (in the same order as they appear in B). Then, it is well-known that

$$L = B^T W B.$$

The above definition makes it obvious that L is a symmetric positive-semidefinite matrix.

Note that given a Laplacian matrix $L \in \mathbb{R}^{n \times n}$ corresponding to an undirected, weighted graph G with m edges and positive edge weights, we can immediately derive B and W . That is, by considering the m non-zero entries L_{ij} with $i < j$, since each such entry corresponds to an edge joining vertices i and j of weight $w_{ij} = -L_{ij}$, we can immediately construct B and W .

1.2 An overview of the problem

Given a Laplacian matrix L corresponding to an underlying graph $G = (V, E)$ with n vertices and m (positively) weighted, undirected edges, consider the following regression problem which was addressed by Spielman and Teng [35, 37, 36].

Problem 1 [LEAST-SQUARES APPROXIMATION WITH LAPLACIAN CONSTRAINTS] *Given as input a Laplacian matrix $L \in \mathbb{R}^{n \times n}$ as described above and a target vector $b \in \mathbb{R}^n$, compute*

$$\arg \min_{x \in \mathbb{R}^n} \|Lx - b\|_2.$$

The minimal ℓ_2 -norm solution vector x_{opt} to the above problem is equal to

$$x_{\text{opt}} = L^\dagger b, \tag{1}$$

where L^\dagger corresponds to the Moore-Penrose generalized inverse.

This formulation is a generalization of the standard problem of solving a system of linear equations of the form $Lx = b$, in order to better handle the rank-deficiency of L . We chose this formulation since our algorithm will make no assumptions on the rank of L . In addition, this formulation will make the comparison with related work on randomized algorithms for matrix problems (see Section 3.3) immediate.

In this setting, Spielman and Teng [36] provided a randomized, relative-error approximation algorithm for Problem 1. The running time of their algorithm is $O(\mathbf{nnz}(A) \log^{c_1} n)$, where $\mathbf{nnz}(A)$ represents the number of non-zero elements of the matrix A , or equivalently the number of edges in the graph G , and c_1 is a small constant. The first step of this algorithm corresponds to performing “spectral graph sparsification,” thereby keeping a small number of edges from G , and thus creating a much sparser Laplacian matrix \tilde{L} . The second step of this algorithm involves using this sparse matrix \tilde{L} as an efficient preconditioner to solve Problem 1 approximately. In order to achieve high precision, this is done in a recursive manner.

While [36] is a major theoretical breakthrough, its applicability is currently hindered by its sheer complexity. In an effort to bridge the gap between theory and practice, recent work of Spielman and Srivastava [34] proposed a much simpler algorithm for the graph sparsification step of [36], by arguing that randomly sampling edges from the graph G with probabilities proportional to the so-called *effective resistances* (see Section 3.1 for definitions) of the edges provides a sparse Laplacian matrix \tilde{L} satisfying the desired properties. On the negative side, in order to approximate the effective resistances of the edges of G efficiently, the Spielman-Srivastava algorithm performs $O(\log n)$ calls to the Spielman-Teng solver, severely hindering its applicability [34]. We should also note that Batson, Spielman, and Srivastava [3] provided a more expensive algorithm for finding even sparser spectral sparsifiers.

Note that the work of Spielman and Teng also addresses a much broader class of matrices, the so-called $SDDM_0$ class, which can be reduced to the Laplacian case. This reduction is described in detail in [10]. For simplicity of presentation, here we will only focus on Laplacian matrices.

1.3 Solving systems of linear equations with Laplacian matrices

Our main result in this paper is a simple algorithm to compute an approximate solution to Problem 1. As with previous algorithms, the first phase will sparsify the input graph, and the second phase will solve the problem on the sparsified graph. Our main algorithm will be described in detail in Section 2. Briefly, in the first phase we will compute a nonuniform sampling probability distribution that depends on the so-called statistical leverage scores [23, 27] associated with the weighted edge-incidence matrix of the input graph. We will then sample a “small” number of edges according to that distribution to construct a sparsified Laplacian matrix \tilde{L} , having $O\left(\frac{n}{\epsilon} \log \frac{n}{\epsilon}\right)$ non-zero entries. Then, in the second phase we will solve the sparsified problem

$$\arg \min_{x \in \mathbb{R}^n} \left\| \tilde{L}x - b \right\|_2 \quad (2)$$

to get the vector $\tilde{x}_{opt} = \tilde{L}^\dagger b$. The resulting vector \tilde{x}_{opt} satisfies (with constant probability)

$$\|x_{opt} - \tilde{x}_{opt}\|_L \leq \epsilon \|x_{opt}\|_L. \quad (3)$$

Recall that the “energy norm” $\|x\|_L$ for any vector $x \in \mathbb{R}^n$ and any matrix $L \in \mathbb{R}^{n \times n}$ is equal to $x^T L x$. Given the sparsified Laplacian \tilde{L} , this second phase will use the conjugate gradient method as a direct solver [39] to solve the sparse least-squares problem of eqn. (2), and thus it will take $O\left(\frac{n^2}{\epsilon} \log \frac{n}{\epsilon}\right)$ time. For dense graphs, this matches the running time of the Spielman-Teng algorithm, while for sparse graphs the Spielman-Teng algorithm is still faster.

The question of computing the statistical leverage scores (either exactly or approximately) is a subtle one, and it is related to the theory-practice disconnect—both for this problem, as well as for other problems to which randomized matrix algorithms have been applied. Thus, we will discuss this topic in greater detail in Section 3.3. Briefly, $O(mn^2)$ time certainly suffices to compute them with standard methods; theoretically, they can be computed in $O(m \log^{c_1} n)$ time, for some small constant c_1 ; and they can be efficiently approximated in the presence of certain resource constraints.

2 An algorithm for solving systems of linear equations

In this section, we will describe our main algorithm to approximate the minimal ℓ_2 -norm solution vector x_{opt} of the least-squares approximation problem with Laplacian constraint matrix (Problem 1). Then, we will state and prove our main quality-of-approximation theorem and discuss the running time of the proposed algorithm.

2.1 Our main algorithm

Algorithm 1 takes as input an $n \times n$ Laplacian matrix L (corresponding to a graph G with n vertices and m positively weighted, undirected edges) and constructs an $n \times n$ sparsified Laplacian matrix \tilde{L} . Finally, it computes the minimal ℓ_2 -norm solution vector \tilde{x}_{opt} of the sparsified problem with a direct solver.

In more detail, the algorithm first computes the edge incidence matrix B and the corresponding diagonal weight matrix W , as described in Section 1.1. Then, it computes a set of probabilities p_1, p_2, \dots, p_m such that the i -th edge of the graph, i.e., the i -th row of the edge-incidence matrix B and the corresponding weight W_{ii} , will be retained with probability proportional to p_i . These probabilities satisfy eqn. (4) and depend on the so-called statistical leverage scores of the matrix $W^{1/2}B$. As we will observe in Section 3.2, these scores are proportional to the effective resistances of the edges of graph G . The parameter β at Step 3 of Algorithm 1 facilitates the use of approximate (as opposed to exact) probabilities and will be further discussed below. It is worth noting that computing the aforementioned probabilities p_i exactly ($\beta = 1$) necessitates $O(mn^2)$ time, which is prohibitive for the proposed application.²

After setting the sparsity parameter r to an appropriate value that guarantees a relative-error approximation to the optimal solution at Step 4, exactly r edges of G are sampled (Step 6) with respect to the computed probabilities. The weights of the retained edges are rescaled (Step 6) and the induced Laplacian \tilde{L} corresponding to the sparsified graph is formed. Note that $\tilde{L} \in \mathbb{R}^{n \times n}$ has at most $n + 2r$ non-zero entries, since its underlying sparsified graph has at most r edges. Then, the sparsified problem

$$\arg \min_{x \in \mathbb{R}^n} \left\| \tilde{L}x - b \right\|_2 \quad (5)$$

is solved in order to return the minimal ℓ_2 -norm solution $\tilde{x}_{opt} = \tilde{L}^\dagger b$. The computational savings emerge since the sparsified problem can be solved efficiently using, for example, conjugate-gradient-type methods as direct solvers. The running time of such methods with input \tilde{L} and b is $O(n(n + 2r))$, where $n + 2r$ is the number of non-zero entries in \tilde{L} .

²Indeed, one goal of this work is to focus further research towards efficient—either provably accurate or heuristic—algorithms to approximate these leverage scores in various settings, thereby leading to faster algorithms for this and related problems.

Input: Laplacian matrix $L \in \mathbb{R}^{n \times n}$, corresponding to a graph G with n vertices and m (positively) weighted edges, $b \in \mathbb{R}^n$, and accuracy parameter $\epsilon \in (0, 1)$.

Output: $\tilde{x}_{opt} \in \mathbb{R}^n$.

1. Compute the edge-incidence matrix $B \in \mathbb{R}^{m \times n}$ and the diagonal edge-weight matrix $W \in \mathbb{R}^{m \times m}$ (see Section 1.1).
2. Let $\Phi = W^{1/2}B \in \mathbb{R}^{m \times n}$.
3. Compute a set of probabilities p_i (for all $i = 1 \dots m$) such that $\sum_{i=1}^m p_i = 1$ and

$$p_i \geq \frac{\beta \left\| (U_\Phi)_{(i)} \right\|_2^2}{\|U_\Phi\|_F^2} \quad (4)$$

for some $\beta \in (0, 1]$. (U_Φ is an orthogonal basis for the column space of Φ and $(U_\Phi)_{(i)}$ is the i -th row of U_Φ .)

4. Set $r = \frac{72c_0^2n}{\beta\epsilon} \log \left(\frac{36c_0^2n}{\beta\epsilon} \right)$, where c_0 is the unspecified constant of Theorem 2.
5. Initialize $S \in \mathbb{R}^{m \times r}$ to be an all-zeros matrix.
6. **For** $t = 1, \dots, r$ **do**
 - Pick $i_t \in 1 \dots m$, where $\text{Prob}(i_t = i) = p_i$;
 - $S_{i_t t} = 1/\sqrt{rp_{i_t}}$;
7. Compute $\tilde{L} = (B^T W^{1/2} S) (S^T W^{1/2} B) \in \mathbb{R}^{n \times n}$.
8. Return $\tilde{x}_{opt} = \tilde{L}^\dagger b$.

Algorithm 1: Approximating the minimal ℓ_2 -norm solution of least-squares problems with Laplacian constraint matrices.

2.2 Approximation accuracy

The following theorem is our main quality-of-approximation result for Algorithm 1.

Theorem 1 *Given Laplacian matrix $L \in \mathbb{R}^{n \times n}$ (corresponding to a graph G with n vertices and m positively weighted edges) and a vector $b \in \mathbb{R}^n$, let $x_{opt} \in \mathbb{R}^n$ be the solution vector of eqn. (1). If $\tilde{x}_{opt} \in \mathbb{R}^n$ is the output of Algorithm 1 for some choice of the accuracy parameter $\epsilon \in (0, 1)$, then, with probability at least $2/3$,*

$$\|x_{opt} - \tilde{x}_{opt}\|_L \leq \epsilon \|x_{opt}\|_L.$$

Proof: By definition, $\|x_{opt} - \tilde{x}_{opt}\|_L = (x_{opt} - \tilde{x}_{opt})^T L (x_{opt} - \tilde{x}_{opt})$. Recall that $L = B^T W B$, where $B \in \mathbb{R}^{m \times n}$ and $W \in \mathbb{R}^{m \times m}$ are the edge-incidence and the diagonal weight matrix respectively (see Section 1.1). Also recall that the diagonal entries of W are positive and thus $W^{1/2}$ is

well-defined. Then,

$$\begin{aligned}
\|x_{opt} - \tilde{x}_{opt}\|_L &= (x_{opt} - \tilde{x}_{opt})^T B^T W B (x_{opt} - \tilde{x}_{opt}) \\
&= \left(W^{1/2} B (x_{opt} - \tilde{x}_{opt}) \right)^T \left(W^{1/2} B (x_{opt} - \tilde{x}_{opt}) \right) \\
&= \left\| W^{1/2} B (x_{opt} - \tilde{x}_{opt}) \right\|_2^2.
\end{aligned} \tag{6}$$

We now use the formulas for x_{opt} and \tilde{x}_{opt} , namely $x_{opt} = L^\dagger b$ (from eqn. (1)) and $\tilde{x}_{opt} = \tilde{L}^\dagger b$. Let $\Phi \in \mathbb{R}^{m \times n}$ denote the matrix $W^{1/2} B$ and let the SVD of Φ be

$$\Phi = U_\Phi \Sigma_\Phi V_\Phi^T. \tag{7}$$

Here $U_\Phi \in \mathbb{R}^{m \times \rho}$, $\Sigma_\Phi \in \mathbb{R}^{\rho \times \rho}$, and $V_\Phi \in \mathbb{R}^{n \times \rho}$, with $\rho \leq n$ being the rank of Φ . Then,

$$L = \Phi^T \Phi = V_\Phi \Sigma_\Phi^2 V_\Phi^T,$$

and thus

$$x_{opt} = L^\dagger b = V_\Phi \Sigma_\Phi^{-2} V_\Phi^T b. \tag{8}$$

Similarly,

$$\tilde{L} = \Phi^T S S^T \Phi = (S^T \Phi)^T (S^T \Phi)$$

and

$$\tilde{x}_{opt} = (S^T \Phi)^\dagger (S^T \Phi)^{\dagger T} b = (S^T U_\Phi \Sigma_\Phi V_\Phi^T)^\dagger (S^T U_\Phi \Sigma_\Phi V_\Phi^T)^{\dagger T} b. \tag{9}$$

Combining eqns. (6), (7), (8), and (9), we get

$$\begin{aligned}
\|x_{opt} - \tilde{x}_{opt}\|_L &= \left\| U_\Phi \Sigma_\Phi V_\Phi^T \left(V_\Phi \Sigma_\Phi^{-2} V_\Phi^T b - (S^T U_\Phi \Sigma_\Phi V_\Phi^T)^\dagger (S^T U_\Phi \Sigma_\Phi V_\Phi^T)^{\dagger T} b \right) \right\|_2^2 \\
&= \left\| \Sigma_\Phi^{-1} V_\Phi^T b - \Sigma_\Phi (S^T U_\Phi \Sigma_\Phi)^\dagger (S^T U_\Phi \Sigma_\Phi)^{\dagger T} V_\Phi^T b \right\|_2^2.
\end{aligned} \tag{10}$$

In the above we used the facts that U_Φ and V_Φ are orthogonal matrices, and $(XV^T)^\dagger = VX^\dagger$ for any orthogonal matrix V . We now employ Theorem 2 of the Appendix in order to argue that $S^T U_\Phi$ is a matrix whose singular values are all close to unity. (This theorem is a variant of a result of Rudelson and Vershynin [31] that was proven as Theorem 4 in the appendix of [24].) More specifically, since $U_\Phi^T U_\Phi = I_\rho$, Theorem 2 argues that with our choice of r at Step 4 of Algorithm 1

$$\mathbf{E} [\|U_\Phi^T S S^T U_\Phi - I_\rho\|_2] \leq \frac{\sqrt{\epsilon}}{6}.$$

Markov's inequality now implies that with probability at least $2/3$

$$\|U_\Phi^T S S^T U_\Phi - I_\rho\|_2 \leq \frac{\sqrt{\epsilon}}{2}. \tag{11}$$

Using standard perturbation theory [38], we get that for all $i = 1, \dots, \rho$,

$$|\sigma_i(U_\Phi^T S S^T U_\Phi) - 1| = |\sigma_i^2(S^T U_\Phi) - 1| \leq \frac{\sqrt{\epsilon}}{2} \tag{12}$$

holds with probability at least $2/3$. (Here $\sigma_i(X)$ denotes the i -th singular value of X .) This implies that the $m \times \rho$ matrix $S^T U_\Phi$ has rank ρ with probability at least $2/3$. The remainder of

the proof will be conditioned on this event holding. Using $(S^T U_\Phi \Sigma_\Phi)^\dagger = \Sigma_\Phi^{-1} (S^T U_\Phi)^\dagger$ (which is only true if $S^T U_\Phi$ has full rank), eqn. (10) becomes

$$\|x_{opt} - \tilde{x}_{opt}\|_L = \left\| \Sigma_\Phi^{-1} V_\Phi^T b - (S^T U_\Phi)^\dagger (S^T U_\Phi)^{\dagger T} \Sigma_\Phi^{-1} V_\Phi^T b \right\|_2^2. \quad (13)$$

We now focus on the matrix $\Omega = S^T U_\Phi \in \mathbb{R}^{m \times \rho}$. Let its SVD be

$$S^T U_\Phi = \Omega = U_\Omega \Sigma_\Omega V_\Omega^T. \quad (14)$$

Since the rank of $S^T U_\Phi$ is ρ , it follows that $U_\Omega \in \mathbb{R}^{m \times \rho}$, $\Sigma_\Omega \in \mathbb{R}^{\rho \times \rho}$, and $V_\Omega \in \mathbb{R}^{\rho \times \rho}$. We now rewrite eqn. (13) using the SVD of Ω :

$$\|x_{opt} - \tilde{x}_{opt}\|_L = \left\| \Sigma_\Phi^{-1} V_\Phi^T b - V_\Omega \Sigma_\Omega^{-2} V_\Omega^T \Sigma_\Phi^{-1} V_\Phi^T b \right\|_2^2. \quad (15)$$

Let $\Sigma_\Omega^{-2} = I_\rho + E$, for some diagonal error matrix E . Using $V_\Omega V_\Omega^T = V_\Omega^T V_\Omega = I_\rho$, eqn. (15) becomes

$$\begin{aligned} \|x_{opt} - \tilde{x}_{opt}\|_L &= \left\| \Sigma_\Phi^{-1} V_\Phi^T b - V_\Omega (I + E) V_\Omega^T \Sigma_\Phi^{-1} V_\Phi^T b \right\|_2^2 \\ &= \left\| V_\Omega E V_\Omega^T \Sigma_\Phi^{-1} V_\Phi^T b \right\|_2^2 \\ &= \left\| E V_\Omega^T \Sigma_\Phi^{-1} V_\Phi^T b \right\|_2^2 \\ &\leq \left\| E V_\Omega^T \right\|_2^2 \left\| \Sigma_\Phi^{-1} V_\Phi^T b \right\|_2^2 \\ &= \|E\|_2^2 \left\| \Sigma_\Phi^{-1} V_\Phi^T b \right\|_2^2. \end{aligned} \quad (16)$$

We now seek to bound the spectral norm of the diagonal matrix E . Notice that the diagonal entries of E satisfy

$$|E_{ii}| = |\sigma_i^{-2}(\Omega) - 1| = |\sigma_i^{-2}(S^T U_\Phi) - 1|.$$

Using the bounds of eqn. (12) we get

$$\begin{aligned} \|E\|_2 &= \max_{i=1 \dots \rho} |\sigma_i^{-2}(S^T U_\Phi) - 1| \\ &= \max_{i=1 \dots \rho} \left| \frac{\sigma_i^2(S^T U_\Phi) - 1}{\sigma_i^2(S^T U_\Phi)} \right| \\ &\leq \frac{\sqrt{\epsilon}/2}{1 - (\sqrt{\epsilon}/2)} \leq \sqrt{\epsilon}. \end{aligned} \quad (17)$$

The last inequality follows since $\epsilon \leq 1$. Combining eqns. (16) and (17), we get

$$\|x_{opt} - \tilde{x}_{opt}\|_L \leq \epsilon \left\| \Sigma_\Phi^{-1} V_\Phi^T b \right\|_2^2. \quad (18)$$

To conclude the proof, notice that using $\Phi = W^{1/2} B$ and eqns. (7) and (8), we get

$$\begin{aligned} \|x_{opt}\|_L &= x_{opt}^T L x_{opt} \\ &= \left(W^{1/2} B x_{opt} \right)^T \left(W^{1/2} B x_{opt} \right) \\ &= \|\Phi x_{opt}\|_2^2 \\ &= \|U_\Phi \Sigma_\Phi V_\Phi^T V_\Phi \Sigma_\Phi^{-2} V_\Phi^T b\|_2^2 \\ &= \left\| \Sigma_\Phi^{-1} V_\Phi^T b \right\|_2^2. \end{aligned} \quad (19)$$

Combining eqns. (18) and (19) concludes the proof of the theorem.

◇

2.3 Running time

We now discuss the running time of Algorithm 1. Steps 1 and 2 are trivial and run in $O(m)$ time. Step 3 necessitates the computation of a probability distribution over the rows of $BW^{1/2}$. Theoretically, this step runs (for $\beta = 1$) in $O(m \log^{c_1} n)$ time, for some small constant c_1 , as described in [3]. (However, in order to achieve this running time it is necessary to perform $O(\log n)$ calls to the Spielman-Teng solver, which essentially renders this computation impractical. Below, we will discuss in more detail several issues related to computing these probabilities in other ways.) Steps 5, 6, and 7 run in $O(m)$ time, since B is a matrix with two non-zero elements per row, W is a diagonal matrix, and the sampling matrix S simply reduces the number of rows in $BW^{1/2}$ from m to r . Finally, at the last step, we invoke a direct solver for the sparse least-squares problem of eqn. (2), which takes $O\left(\frac{n^2}{\epsilon} \log \frac{n}{\epsilon}\right)$ time. Thus, from a theoretical perspective, using the fact that $m \leq n^2$, the running time Algorithm 1 is $O\left(\frac{n^2}{\epsilon} \left(\log \frac{n}{\epsilon}\right) (\log^{c_1} n)\right)$.

3 Connecting graph resistances and statistical leverage scores

In this section, we will show that the effective resistances of the edges of a graph G with n vertices and m positively weighted undirected edges are proportional to the statistical leverage scores of the rows of the matrix $W^{1/2}B$ (recall our definitions in Section 1.1). Although this connection is straightforward from technical perspective, it is of considerable interest due to the insights it provides.

3.1 Review of effective resistance and statistical leverage

We start with the following definition of the *effective resistance* of an edge of a graph:

Definition 1 *Given $G = (V, E)$, a connected, weighted, undirected graph with n nodes, m edges, and corresponding edge weights $w_e \geq 0$, for all $e \in E$, let*

$$L = B^T W B \tag{20}$$

denote the $n \times n$ Laplacian matrix of G (see Section 1.1 for notation). The effective resistances R_e across all edges $e \in E$ are given by the diagonal entries of the matrix

$$R = B L^\dagger B^T, \tag{21}$$

where L^\dagger denotes the Moore-Penrose generalized inverse of L .

Clearly, from standard matrix algebra, the effective resistances of all the edges of G can be computed in $O(n^3)$ time. Moreover, if we let G denote an electrical network, in which each edge $e \in E$ corresponds to a resistor of resistance $1/w_e$, then the effective resistance R_e between two vertices can be defined as the potential difference induced between the two vertices when a unit of current is injected at one vertex and extracted at the other vertex. Finally, effective resistances have a wide range of applications, including not only theoretical applications such as analyzing diffusion processes and random walks on graphs, but also very practical applications such as analyzing clustering and community structure in large informatics networks.

A seemingly-unrelated notion is that of the *statistical leverage scores* of the rows of a matrix:

Definition 2 Given an $m \times n$ matrix A , with $m > n$, the statistical leverage scores of the rows of A are the m diagonal elements of the projection matrix onto the span of the columns of A . That is, if the matrix U_A denotes any orthogonal basis for the column space of A , then the diagonal elements of the projection matrix P_A onto the span of those columns are given by

$$(P_A)_{ii} = (U_A U_A^T)_{ii} = \|(U_A)_{(i)}\|_2^2,$$

where $(U_A)_{(i)}$ denotes the i -th row of the matrix U_A .

Clearly, all the statistical leverage scores can be computed in $O(mn^2)$ time. Note that these scores could be defined for any $m \times n$ matrix A with $m \leq n$. In that case, however, if A is not rank-deficient, then all the scores are trivially equal to unity. Importantly, the statistical leverage scores have a natural interpretation in terms of “importance” or “influence” or “leverage” of the corresponding constraint/row of A in the overconstrained least squares optimization problem $\min_x \|Ax - b\|_2$. As such, they have been of interest historically in diagnostic regression analysis [15].

More generally, given a rank parameter k , one can define the *statistical leverage scores relative to the best rank- k approximation to A* to be the m diagonal elements of the projection matrix onto the span of the best rank- k approximation to A . These generalized scores have been used recently as importance sampling probabilities to obtain relative-error approximation algorithms for regression [22, 24], and they were essential for the extension of these ideas to relative-error low-rank matrix approximation [23, 27] problems. Prior work [14, 2] has also used term *incoherent* to refer to the situation when no leverage score is particularly large.

3.2 A simple lemma

We now describe a connection between graph resistances and statistical leverage scores. Although this connection is not so surprising from a technical perspective—indeed, it is obvious once it is pointed out—it is useful for the insights it provides.

Lemma 1 Let the matrix $\Phi = W^{1/2}B \in \mathbb{R}^{m \times n}$ denote the edge-incidence matrix of a graph G rescaled by $W^{1/2}$. The statistical leverage scores associated with Φ are (up to scaling) equal to the effective resistances of all edges of a weighted graph G . That is, if ℓ_i is the leverage score associated with the i -th row of Φ , then ℓ_i/w_i is the effective resistance of the i -th edge.

Proof: Consider the matrix

$$P = W^{1/2}B(B^T W B)^+ B^T W^{1/2} \in \mathbb{R}^{m \times m},$$

and notice that $P = W^{1/2}R W^{1/2}$ is simply a rescaled version of the $m \times m$ matrix $R = B L^+ B^T$, whose diagonal entries are exactly equal to the effective resistances of all the edges of G . Since $\Phi = W^{1/2}B$, it follows that

$$P = \Phi(\Phi^T \Phi)^+ \Phi^T.$$

Let U_Φ denote an orthogonal matrix spanning the column space of Φ . Then $P = U_\Phi U_\Phi^T$, from which it follows that the diagonal elements of P are equal to

$$P_{ii} = (U_\Phi U_\Phi^T)_{ii} = \|(U_\Phi)_{(i)}\|_2^2.$$

This concludes the proof of the lemma. ◇

3.3 Usefulness of statistical leverage in randomized matrix algorithms

The connection between statistical leverage and effective resistance is of interest in attempts to make nearly-linear-time linear equation solvers more practical. The reason is that statistical leverage has proven to be the key structural quantity to understand in order to bridge the “theory-practice gap” between theoretical work on randomized algorithms for large matrices, and applications (both numerical-implementation and data-analysis applications) of this “randomized matrix algorithm” paradigm [28, 27, 24, 30, 2]. In this section, we review some of the “lessons learned,” in the hope that they provide insights on how to bridge the theory-practice gap for solving linear equations defined by a Laplacian constraint matrices.

Recall that much work, including, *e.g.*, our previous work [19, 20, 21], followed that of Frieze, Kannan, and Vempala [25], in which columns and/or rows from a matrix A are randomly sampled according to a probability distribution that depends on the Euclidean norms of those columns/rows. In this case, worst-case additive-error guarantees of the form

$$\|A - P_{C,k}A\|_F \leq \|A - A_k\|_F + \epsilon \|A\|_F \quad (22)$$

can be obtained, with high probability.³ Although these algorithms were motivated by resource-constrained computational environments, they have several drawbacks with respect to numerical applications and data analysis applications more generally. First, worst-case additive-error bounds are quite coarse. Second, these algorithms were not immediately-relevant to common problems, as they are typically formulated, in scientific computing and numerical linear algebra. Third, the insights provided by the sampling probabilities into the data are limited—the probabilities are often uniform due to data preprocessing, or they may correspond, *e.g.*, simply to the degree of a node if the data matrix is derived from a graph.

Importantly, each of these three problems was solved by the introduction of importance sampling probabilities that depend on the statistical leverage scores.⁴

- First, by using importance sampling probabilities that depend on the leverage scores, it was shown [23, 27] that one could randomly sample a “small” number of columns to obtain worst-case relative-error guarantees of the form

$$\|A - P_{C,k}A\|_F \leq (1 + \epsilon) \|A - A_k\|_F, \quad (23)$$

with high probability.

- Second, algorithms that were comparable to or better than previously-existing algorithms were provided for the following two very traditional scientific computing problems:
 - **Overconstrained Least Squares.** Let A be an $m \times n$ matrix A , with $m \gg n$, and consider solving $x_{opt} = \arg \min_x \|Ax - b\|_2$. In previous work [22, 23, 24], we proposed a simple, sampling-based, algorithm for solving this problem: first, compute the statistical leverage scores of the rows of A ; then, use these scores to construct an importance sampling probability distribution to sample a “small” number of rows of A and the corresponding elements of b ; and finally, solve the induced, much smaller but still overconstrained, regression problem using only those (suitably rescaled) rows

³Here $P_{C,k}A$ denotes the projection of A on a rank- k subspace spanned by the columns of C .

⁴Although these probabilities were introduced in [22, 23] and were used in solving two very traditional numerical linear algebra problems in [24, 12], the connection with leverage scores wasn’t made explicit until [27].

of A and the corresponding elements of b . Strong relative error guarantees for this overconstrained⁵ regression problem were proven with this approach [22, 23].

- **Column Subset Selection Problem.** Let A be an $m \times n$ matrix, and let k be a positive integer. Then, pick k columns of A forming an $m \times k$ matrix C such that the residual $\|A - P_C A\|_\xi$, where $\xi = 2$ or F denotes the spectral norm or Frobenius norm, is minimized over all possible $\binom{n}{k}$ choices for the matrix C . Previously [12, 11], we developed a two-phase algorithm that uses the nonuniformity structure defined by the statistical leverage scores in an essential way to provide theoretical and empirical results for both the spectral and Frobenius norm that were competitive or better than previously existing results.
- Third, the insights into the matrix provided by statistical leverage scores (in both numerical and data applications) can be quite refined. The insights are used in very different ways, depending on whether one is interested in high-quality numerical implementations or large-scale data analysis applications.
 - **Numerical Implementation Applications.** Here, one wants to provide fast high-quality numerical implementations, and one is typically interested in the error parameter to be very small, *e.g.*, $\epsilon \approx 10^{-16}$. For example, with respect to the overconstrained least-squares regression problem, performing an exact computation of the statistical leverage scores of the rows of A is no faster than exactly solving the original regression problem. Sarlós [33, 24] addressed this problem by preprocessing the matrix A and the vector b with the randomized Hadamard transform of Ailon and Chazelle [1]. This preprocessing step made the statistical leverage scores almost uniform—effectively “washing out” any nonuniformities defined by the leverage scores, thereby densifying the matrix if it was sparse—thus leading to the first randomized, relative-error algorithm for least-squares problems that runs asymptotically faster than $\Theta(mn^2)$ time. High-quality implementations of such algorithms have appeared [30, 2], and they highlight the significant practical applicability of this approach.
 - **Data Analysis Applications.** Here, one may want $\epsilon \approx 0.1$, and one is typically interested in obtaining insight with respect to some downstream data analysis goal. In such cases, SVD-based methods are often chosen for computational convenience, rather than because the statistical assumptions underlying their use are satisfied by the data—a fact which means that the leverage scores are often extremely nonuniform in a way that correlates strongly with what practitioners know about the data [28, 27, 11, 13] problems. Thus, far from “washing out” this nonuniformity structure, one is interested in identifying and exploiting it. Intuitively, conditioned on being reliable, more “outlier-like” data points may be the most important and informative.

This brings us to the question of how to compute these statistical leverage scores, or equivalently the effective resistances, which is an issue that gets to the heart of the theory-practice gap. Depending on the application and the resource constraints, there are several alternatives:

⁵Note that it is easy to show that similar results hold for the very underconstrained problem. Let A be an $m \times n$ matrix, with $m \ll n$, and consider the problem of finding the minimum-length solution to $x_{opt} = \operatorname{argmin}_x \|Ax - b\|_2 = A^+ b$. Sampling variables or columns from A can be represented by postmultiplying A by a $n \times c$ (with $c > m$) column-sampling matrix S to construct the (still underconstrained) least-squares problem: $\tilde{x}_{opt} = \operatorname{argmin}_x \|ASS^T x - b\|_2 = A^T (AS)^{T+} (AS)^+ b$. The second equality follows by inserting $P_{A^T} = A^T A^{T+}$ to obtain $ASS^T A^T A^{T+} x - b$ inside the $\|\cdot\|_2$ and recalling that $A^+ = A^T A^{T+} A^+$ for the Moore-Penrose pseudoinverse. If one randomly samples $c = O((n/\epsilon^2) \log(n/\epsilon))$ columns according to “column-leverage-score” probabilities, *i.e.*, the diagonal elements of the projection matrix onto the row space, then it can be proven that $\|x_{opt} - \tilde{x}_{opt}\|_2 \leq \epsilon \|x_{opt}\|_2$ holds, with high probability.

- Compute the scores by calling the Spielman-Teng nearly-linear time solver. This algorithm runs in $O(\text{nnz}(A) \log^{c_1} n)$ time, where $\text{nnz}(A)$ represents the number of non-zero elements of the matrix A , or equivalently the number of edges in the graph G , and c_1 is a small constant. This method works for computing the leverage scores of Laplacian matrices; and in this case it is, theoretically, the best method.
- Compute the scores by computing an “exact” basis for the column space of the $m \times n$ matrix $\Phi = W^{1/2}B$. This takes $O(mn^2)$ time and works for general matrices. For Laplacian matrices it is clearly expensive, given that the weighted edge-incidence matrix is very sparse.
- Compute an approximation to the scores based on iterative sampling and volume sampling ideas that have been used in relative-error low-rank matrix approximations [18, 17]. This might be of interest if a pass-efficient model is an appropriate model for data access.
- Compute an approximation to the scores based on numerical methods to, e.g., compute an estimator for the diagonal of a matrix [5]. These numerical methods are particularly appropriate for large matrices when matrix-vector products are easy to evaluate; they have proven useful in uncertainty quantification [4]; and they draw on the observation that the leverage scores, being proportional to the diagonal elements of a projection matrix, have a natural interpretation in scientific computing in terms of density matrices and Green’s functions [32].

These alternate approaches are of particular interest since data points with high leverage scores often have natural interpretations in terms of processes generating the data matrices [27]. Moreover, an examination of the details of these methods illustrates that problems are parameterized within theoretical computer science in very different ways than they are parameterized in scientific computing. Finally, an important issue to keep in mind is that in most applications, one does not need a uniformly good approximation to all the leverage scores, but instead one needs a good approximation only to the “high leverage” data points.

4 Conclusion

Several open problems suggest themselves. On the theoretical side: Can one draw on the original ideas of Spielman and Teng in order to develop an algorithm with the simplicity of ours and with the running time approximation of theirs? Similarly, can we get the $O(n \log n)$ factor, which currently is due to the result of Rudelson and Vershynin [31], down to $O(n)$, even for some classes of graphs, thereby obtaining a more immediately practical version of the result of Batson, Spielman, and Srivastava [3]? On the more applied side: How rapidly can we approximate (even with a one-sided approximation) the statistical leverage scores, either for general $m \times n$ matrices A and arbitrary rank parameter k , or under some realistic generative model? Similarly, can one use the connection between statistical leverage and effective resistance to design improved heuristics, given knowledge about the processes generating the data?

We conclude by noting that the last two questions are of particular interest. Although much of the recent work on using Laplacian preconditioners has focused on nearly-linear-time solvers for computing “exact” solutions, *i.e.*, with the error parameter ϵ set to machine precision, there are many other applications of these ideas. For example, in machine learning, Ravikumar and Lafferty used preconditioner approximations for doing approximate inference in probabilistic graphical models [29]. This connection should not be surprising, as much of the work on the “randomized algorithms for matrices” paradigm has been motivated by large-scale data applications. In many of these data analysis applications, however, not only is setting $\epsilon = 10^{-16}$ not of interest, doing

so would actually lead to “worse” answers than setting it, say, as $\epsilon = 0.1$. If other recent applications of the randomized algorithms paradigm are any guide [30, 2, 27, 6], then the issues that will arise when thinking of ϵ as extremely small and trying to couple newer randomized algorithmic methods with traditional numerical methods [30, 2] will be very different than the issues that arise in applications where the data are much less well-structured and much-coarser ϵ ’s are of interest [27, 6].

References

- [1] N. Ailon and B. Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, pages 557–563, 2006.
- [2] H. Avron, P. Maymounkov, and S. Toledo. Blendenpik: Supercharging LAPACK’s least-squares solver. Manuscript. (2009).
- [3] J. Batson, D.A. Spielman, and N. Srivastava. Twice-Ramanujan sparsifiers. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, pages 000–000, 2009.
- [4] C. Bekas, A. Curioni, and I. Fedulova. Low cost high performance uncertainty quantification. In *Proceedings of the 2nd Workshop on High Performance Computational Finance*, page Article No.: 8, 2009.
- [5] C. Bekas, E. Kokiopoulou, and Y. Saad. An estimator for the diagonal of a matrix. *Applied Numerical Mathematics*, 57:1214–1229, 2007.
- [6] M.-A. Belabbas and P. J. Wolfe. Spectral methods in machine learning and new strategies for very large datasets. *Proc. Natl. Acad. Sci. USA*, 106:369–374, 2009.
- [7] M. Bern, J.R. Gilbert, B. Hendrickson, N. Nguyen, and S. Toledo. Support-graph preconditioners. *SIAM Journal on Matrix Analysis and Applications*, 27(4):930–951, 2006.
- [8] E.G. Boman, D. Chen, B. Hendrickson, and S. Toledo. Maximum-weight-basis preconditioners. *Numerical Linear Algebra with Applications*, 11(8-9):695–721, 2004.
- [9] E.G. Boman and B. Hendrickson. Support theory for preconditioning. *SIAM Journal on Matrix Analysis and Applications*, 25(3):694–717, 2003.
- [10] E.G. Boman, B. Hendrickson, and S. Vavasis. Solving elliptic finite element systems in near-linear time with support preconditioners. *SIAM Journal on Numerical Analysis*, 46(6):3264–3284, 2008.
- [11] C. Boutsidis, M.W. Mahoney, and P. Drineas. Unsupervised feature selection for principal components analysis. In *Proceedings of the 14th Annual ACM SIGKDD Conference*, pages 61–69, 2008.
- [12] C. Boutsidis, M.W. Mahoney, and P. Drineas. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the 20th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 968–977, 2009.
- [13] C. Boutsidis, M.W. Mahoney, and P. Drineas. Unsupervised feature selection for the k -means clustering problem. In *Annual Advances in Neural Information Processing Systems 22: Proceedings of the 2009 Conference*, 2009.

- [14] E. Candes and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23(3):969–985, 2007.
- [15] S. Chatterjee and A.S. Hadi. *Sensitivity Analysis in Linear Regression*. John Wiley & Sons, New York, 1988.
- [16] F.R.K. Chung. *Spectral graph theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, 1997.
- [17] A. Deshpande and L. Rademacher. Efficient volume sampling for row/column subset selection. Technical report. Preprint: arXiv:1004.4057 (2010).
- [18] A. Deshpande, L. Rademacher, S. Vempala, and G. Wang. Matrix approximation and projective clustering via volume sampling. *Theory of Computing*, 2:225–0247, 2006.
- [19] P. Drineas, R. Kannan, and M.W. Mahoney. Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication. *SIAM Journal on Computing*, 36:132–157, 2006.
- [20] P. Drineas, R. Kannan, and M.W. Mahoney. Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix. *SIAM Journal on Computing*, 36:158–183, 2006.
- [21] P. Drineas, R. Kannan, and M.W. Mahoney. Fast Monte Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition. *SIAM Journal on Computing*, 36:184–206, 2006.
- [22] P. Drineas, M.W. Mahoney, and S. Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1127–1136, 2006.
- [23] P. Drineas, M.W. Mahoney, and S. Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30:844–881, 2008.
- [24] P. Drineas, M.W. Mahoney, S. Muthukrishnan, and T. Sarlós. Faster least squares approximation. Technical report. Preprint: arXiv:0710.1435v3 (2007).
- [25] A. Frieze, R. Kannan, and S. Vempala. Fast Monte-Carlo algorithms for finding low-rank approximations. *Journal of the ACM*, 51(6):1025–1041, 2004.
- [26] G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 1996.
- [27] M.W. Mahoney and P. Drineas. CUR matrix decompositions for improved data analysis. *Proc. Natl. Acad. Sci. USA*, 106:697–702, 2009.
- [28] P. Paschou, E. Ziv, E.G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M.W. Mahoney, and P. Drineas. PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genetics*, 3:1672–1686, 2007.
- [29] P. Ravikumar and J. Lafferty. Preconditioner approximations for probabilistic graphical models. In *Annual Advances in Neural Information Processing Systems 18: Proceedings of the 2005 Conference*, 2006.
- [30] V. Rokhlin and M. Tygert. A fast randomized algorithm for overdetermined linear least-squares regression. *Proc. Natl. Acad. Sci. USA*, 105(36):13212–13217, 2008.

- [31] M. Rudelson and R. Vershynin. Sampling from large matrices: an approach through geometric functional analysis. *Journal of the ACM*, 54(4):Article 21, 2007.
- [32] Y. Saad, J. R. Chelikowsky, and S. M. Shontz. Numerical methods for electronic structure calculations of materials. *SIAM Review*, 52(1):3–54, 2010.
- [33] T. Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 143–152, 2006.
- [34] D.A. Spielman and N. Srivastava. Graph sparsification by effective resistances. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pages 563–568, 2008.
- [35] D.A. Spielman and S.-H. Teng. A local clustering algorithm for massive graphs and its application to nearly-linear time graph partitioning. Technical report. Preprint: arXiv:0809.3232 (2008).
- [36] D.A. Spielman and S.-H. Teng. Nearly-linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. Technical report. Preprint: arXiv:cs/0607105 (2006).
- [37] D.A. Spielman and S.-H. Teng. Spectral sparsification of graphs. Technical report. Preprint: arXiv:0808.4134 (2008).
- [38] G.W. Stewart and J.G. Sun. *Matrix Perturbation Theory*. Academic Press, New York, 1990.
- [39] L.N. Trefethen and D. Bau III. *Numerical Linear Algebra*. SIAM, Philadelphia, 1997.

Appendix

Let $A \in \mathbb{R}^{m \times n}$ be any matrix. Consider the following algorithm, which is essentially the algorithm in page 876 of [23]. This algorithm constructs a matrix $C \in \mathbb{R}^{m \times c}$ consisting of c sampled and rescaled columns of A .

Data : $A \in \mathbb{R}^{m \times n}$, $p_i \geq 0, i \in [n]$ s.t. $\sum_{i \in [n]} p_i = 1$, positive integer $c \leq n$.

Result : $C \in \mathbb{R}^{m \times c}$

Initialize $S \in \mathbb{R}^{m \times c}$ to be an all-zero matrix.

for $t = 1, \dots, c$ **do**

Pick $i_t \in [n]$, where $\mathbf{Prob}(i_t = i) = p_i$;

$S_{i_t t} = 1/\sqrt{cp_{i_t}}$;

end

Return $C = AS$;

Algorithm 2: The EXACTLY(c) algorithm.

Next, we state a theorem that provides a bound for the approximation error $\|AA^T - CC^T\|_2$. We used this in the proof of our main theorem in Section 2 in order to argue that the singular values of the “sampled orthogonal” matrix $S^T U_\Phi$ are all close to unity. In this form, the theorem was proven as Theorem 4 in the Appendix of [24], but it is a variant of the well-known result of Rudelson and Vershynin [31].

Theorem 2 *Let $A \in \mathbb{R}^{m \times n}$ with $\|A\|_2 \leq 1$. Construct C using the EXACTLY(c) algorithm and let the sampling probabilities p_i satisfy*

$$p_i \geq \beta \frac{\|A^{(i)}\|_2^2}{\|A\|_F^2} \quad (24)$$

for all $i \in [n]$ for some constant $\beta \in (0, 1]$. Let $\epsilon \in (0, 1)$ be an accuracy parameter, assume $c_0^2 \|A\|_F^2 \geq 4\beta\epsilon^2$, and let

$$c = 2 \left(\frac{c_0^2 \|A\|_F^2}{\beta\epsilon^2} \right) \log \left(\frac{c_0^2 \|A\|_F^2}{\beta\epsilon^2} \right).$$

(Here c_0 is the unknown constant of Theorem 3.1, p. 8 of [31].) Then,

$$\mathbf{E} [\|AA^T - CC^T\|_2] \leq \epsilon.$$

Finally, it is worth noting that the condition $c_0^2 \|A\|_F^2 \geq 4\beta\epsilon^2$ is trivially satisfied for any matrix A such that $\|A\|_F^2 \geq 4$ assuming $c_0 \geq 1$.